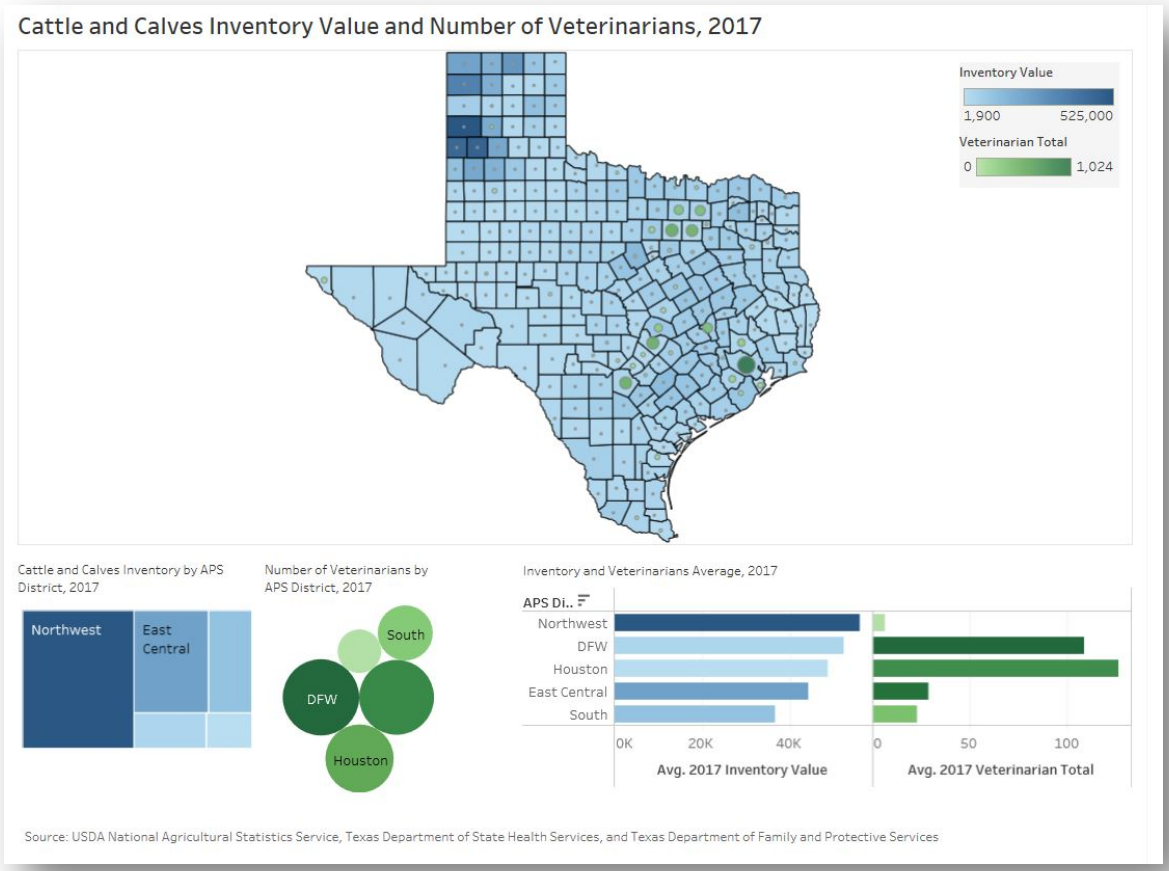
Business Intelligence – Final Project Report

TTU – Rawls College of Business
Master in Data Science
Professor: Dr. Lucus
Student: Luciano Vilas Boas

Brief Summary

Texas has the most cattle inventory in the United States and accounts for roughly 13% of that inventory in America. Texas is also on the top when comes to the annual mean wage for veterinarians; Texas is the state that paid them more. Yet, the state faces a discrepancy and shortage of these professionals to take care of its enormous cattle population. The problem gets even worse due the fact that most veterinarians are located far away from where the vast majority of cattle are located.

Currently, Texas is seeking to open a School of Veterinary Medicine and this work aims to provide some basic visualization of this problem and help decision makers during the process. Below is a sample of the dashboard which pictures the situation:



Analysis of Data

◦ **What questions can these data potentially answer?**

The merged data can potentially answer questions such as, for example: which counties has more cattle and calves inventory and veterinarians. It can also shows us these numbers by regions.

◦ **What are the potential valuable data items exists within the data?**

We can find the rank of counties with more vets, and we can also see the rate of vets per 100,000 population – this information is available in the merged file, but it's not being displayed in the dashboard. All information and data is given at the county level, which gives us a realistic picture of what we are tackling.

◦ **How might they be applied for direct business application and indirect business applications?**

For direct business applications, the dashboard created on Tableau can potentially help on getting funding to bring the School of Veterinary to West Texas. Indirectly, can help people decide whether to pursue a career as veterinarian in Texas or not.

◦ **What do you suggest as potential usages for different variables within the dataset?**

The dataset can be used separately, however, I believe it's more powerful and brings more value when looked merged. The user can access the vets information separately to find where most of them are and where we have shortage of it. In that same note, you can use the Cattle and Calves Inventory to figure out where they are located across the state.

Data Cleaning

◦ **What is the overall quality of the data?**

The overall quality was good. But I had to work on some tweaks as following:

The data about veterinarians (file: veterinarians_2017.csv) has this following character "-" which were causing me trouble when merging, so I had to replace this character to zero, which did not changed the original data too much.

Regarding the data about the Texas regions (file: texas_regions.csv), I had to fix the name of one county in order to proper join the 3 different datasets. I had to change "DeWitt" to "De Witt".

The data about Cattle and Calves Inventory did not require me to clean it, especially because the columns I wanted for the join were already in a good shape. I use python to download this file straight to csv and I did not clean it using Python; but I got rid of unnecessary columns using Pentaho.

◦ **What variables contained missing data?**

The veterinarians' data had that character ("-") which can be seen as missing data in some way. The reason for that is because some counties in Texas do not have any veterinarians. So the variables that carry out calculations based on number of vets were marked with the "-". These variables are: "Ratio of 2017 Population to Veterinarian", and "Rank".

Cattle and Calves Inventory has some missing values, however, those values are no relevant for merging and reporting. Some of those missing variables are "Week Ending", and "Zip Code". As mentioned previously, I disregarded these columns on Pentaho.

Texas Regions did not have any missing value.

◦ **What kinds of missing value exists in the dataset and which variables are they related to?**

The missing values were mostly numeric. The previous answer already covered that.

◦ **What methods did you use to clean the missing data?**

In Python, the code I used to clean the missing data was "replace". Basically, I replaced the "-" for zeros.

I also use Pentaho to "clean" the missing data, especially on the Cattle and Calves Inventory. All I did was getting riding of the variables I did not want to use or was not necessary for my report (example are zip code and week ending columns).

Data Merging

◦ **What were the common elements between both datasets?**

All three datasets I used have in common the "County" variable as the name to be used to join. The best scenario would be having the county id number, which is a unique code that every single county has. By using that, we mitigate the trouble I had when merging those three files, which was the fact that one county name (De Witt) had a different spelling in one of the datasets, which caused me an issue. Unfortunately, the datasets I picked did not have that coding, but I was aware of this potential problem and I was expecting to come across this situation. My solution was changing the name of the mentioned county to match with the other datasets and enable the join.

◦ **Were there any issues with multilevel measurement in the final dataset?**

Apart from the issue with the county name that I have already explained, the merge data was very accurate and matching with its independent files. I have not spotted any multilevel measurement in the final dataset.

◦ **What variables are more valuable combined than being in separate datasets?**

Combining number of vets and cattle and calves inventory has definitely more value together than separate. As shown in the dashboard, the counties with higher inventory are those with few number of vets. In the other hand, the counties with more vets are not those that would have more inventory. Additionally, showing these two numbers by regions also illustrates that in a different way and empathizes the discrepancy we have.

◦ **In what ways has the data become more valuable since being merged? i.e. what new business insights can be generated due to the combined datasets rather than the sets being separate**

Some of the major insight is the fact that the merged dataset illustrates the need for more attention in West Texas when comes to give more access to care and treatment, as well as growing the workforce, in that area. That could be an opportunity for new businesses and professionals that want to take advantage of the need those counties have for vets.

Analysis of Visualizations

◦ **How well does your visualization adhere to the principles and characteristics of a good visualization?**

The visualization was made to be simply understood for anybody that wants to see the contrast between where the cattle and veterinarians are located in Texas at the county level. In that sense, I used blue for cattle and green for veterinarians. I repeated these colors throughout the dashboard as an attempt to create this association: blue = cattle and green = vets. I also worked on creating a gestalt approach that emphasizes a pattern: West Texas = more cattle and less vets, and the rest of Texas = less cattle and more vets.

Additionally, the counties in the map are in blue and most counties has a square shape, so I tried to link the counties in the map with the "Cattle Inventory by APS Districts" heat chart, which also has square shapes. I also did the same thing with the vets data: in the map we see the green circles for the vets and I also have a bubble chart that shows the number of vets by APS districts. I am associating colors and shapes as an effort to create a "no brainer" association.

I summary, the visualization tackles and adheres the principles of a good visualization.

◦ **How well does your visualization adhere to the concept of natural processing? Are there things in your graph that are necessary but do not have a natural processing correlate?**

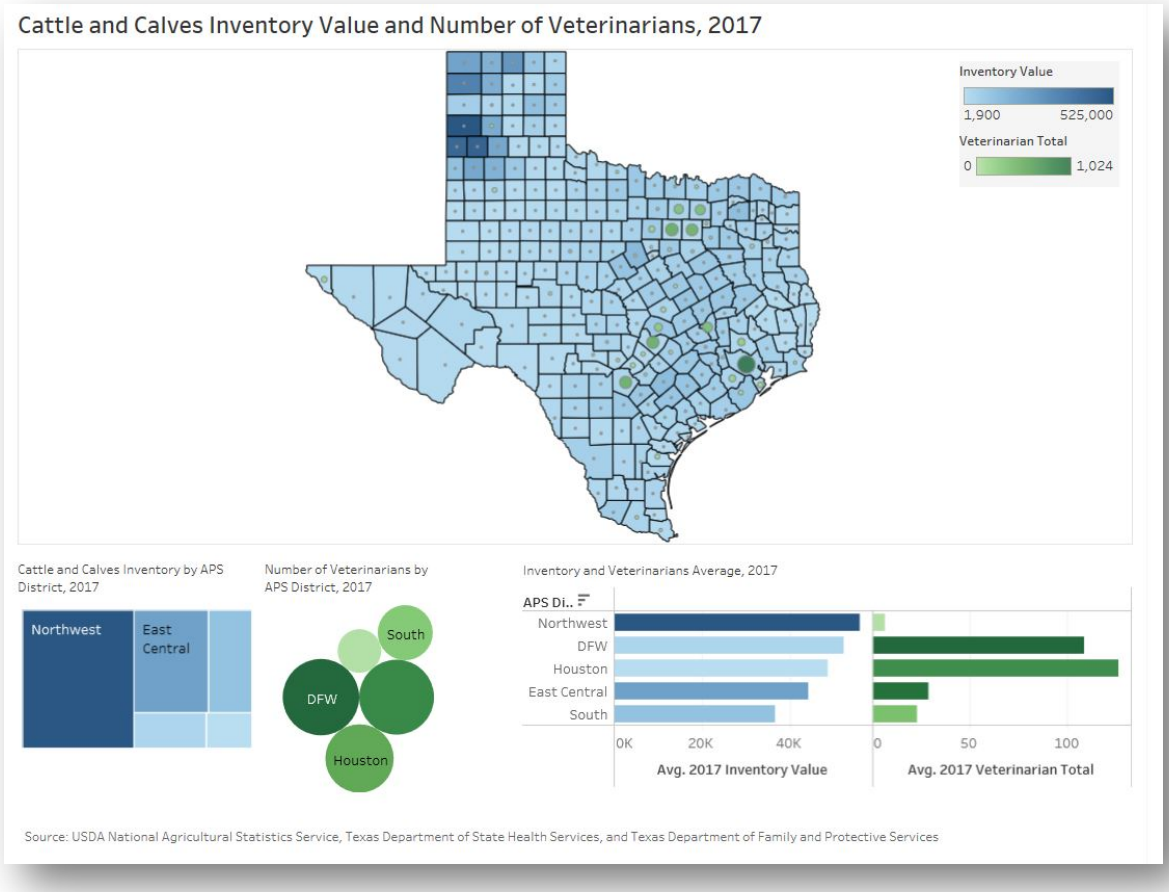
The fact that I focused on putting Texas's map in the center of the dashboard has the purpose of creating this natural effect of familiarization with the map. The audience does not have to think too much to figure out that we are focusing on Texas (I even made sure to let the background in white instead of using the divisions and information of other states as usual on Tableau maps). I also used a bar chart which is a very common way to show data.

Maybe, depending on the viewer's prior knowledge, the heat and bubble chart might look unnatural, but I believe they play and important rule on making the associations that I've already mentioned previously.

All and all, I think the dashboard does a good job conveying the concept of natural association because the visualization appeals to represent the data in a way that is similar to the user's environment.

◦ **Copies of your visualizations.**

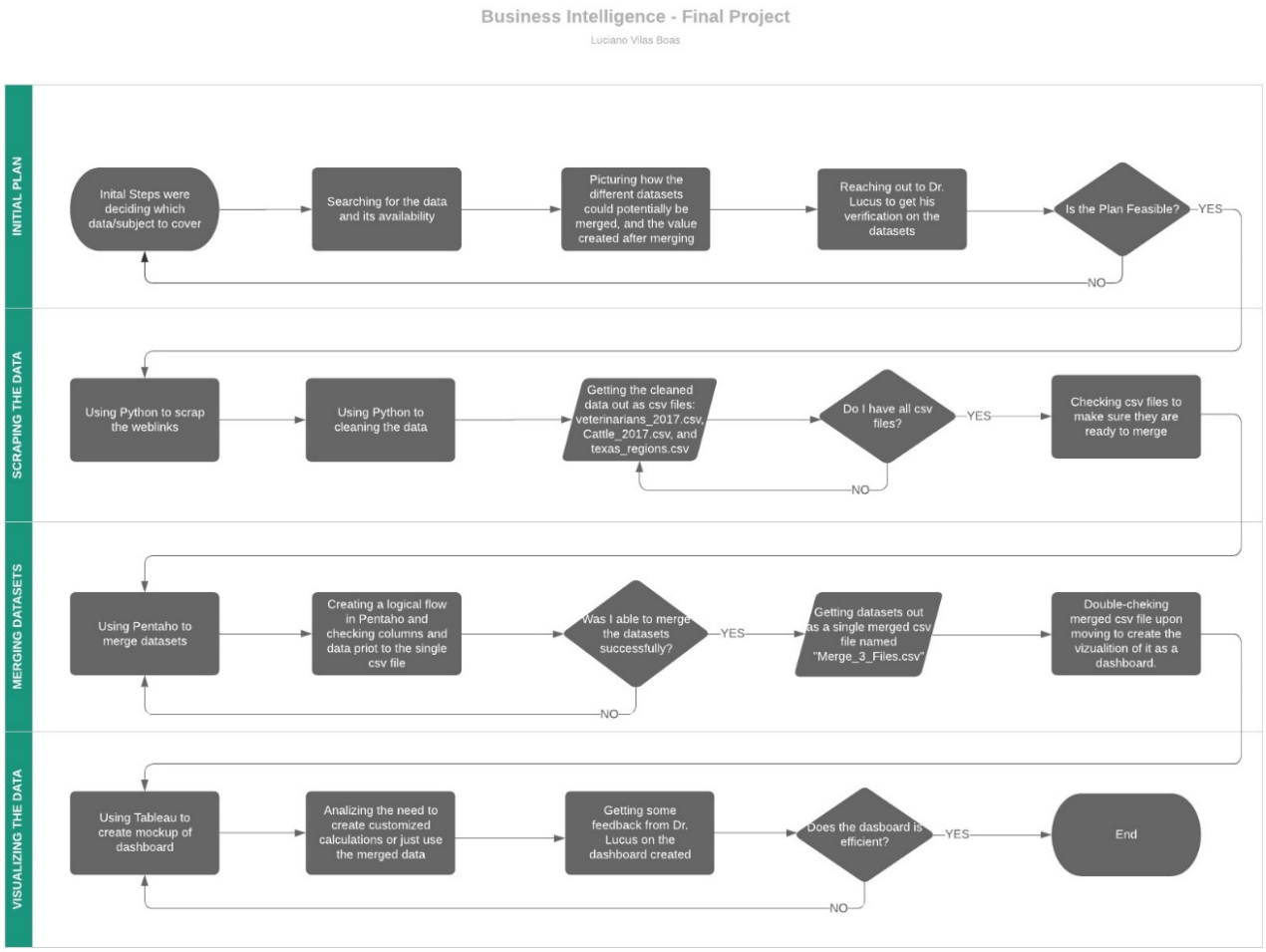
Here is a copy of my visualization:



Flow diagram of your project.

◦ **This should illustrate all steps necessary to gather the data to visualizing the data. It should also illustrate which "files" and "language" are being used at each step.**

Here is my flow chart:



Instructions for code

◦ **Note, this should be explicit step-by-step instructions, including any variables that might need to be manipulated.**

Working on Python:

Here are the instructions for dealing with the Python files (in no particular order):

1. The Python file named "**Texas_Regions.py**" gets information on Texas regions divided by APS Districts, for example. Below is the step-by-step to generate a csv file with this intel:
   1.1. First you need to call out the Python libraries in the file;
   1.2. Second you'll have to access the url and be able to see the data printed in Python console. In order to do that it's required to run all those codes in the second step together (select all and run);
   1.3. Third step is to run the codes to check if there is something wrong with the data or if it needs to be cleaned.
   1.4. Fourth action is to change the name of the county "DeWitt" to "De Witt". For that we are using the "loc" function which allows us to indicate the index and columns we are changing. We need to change the name so we can match it when joining with the other files.
   1.5. Last step is printing out the csv file with this change. This csv file is "texas_regions.csv".

2. The Python file "**VETERINARIANS_2017.py**" has data related to the number of veterinarians in Texas by county.
   2.1. First you need to call out the Python libraries in the file;
   2.2. Second you'll have to access the url and be able to see the data printed in Python console. In order to do that it's required to run all those codes in the second step together (select all and run);
   2.3. Third step is to run the codes to check if there is something wrong with the data or if it needs to be cleaned.
   2.4. Forth action is just see how many wrong character, which is "-", we have in order to get an idea of how to fix the issue;
   2.5. Fifth step is to replace the wrong character from "-" to "zeros".
   2.6. Last action is to print out the csv file with the correction we just did. This csv file is "veterinarians.csv".

3. The Last Python file is the "**Cattle_2017.py**". This file has intel on cattle and calves inventory at the county level. The source that holds this data works with a dynamic table, which has the data we are looking to scrap. I've tried many different way to do that, which can be challenging, and the easiest way it using Python to access the website and download the data straight into a csv files. Potential cleaning and changing in data would be done using Pentaho, if needed.
   3.1. First step is call the libraries;
   3.2. Second and last action is to run Chromedriver to access and download the data as csv.

After creating these three csv files, we will use Pentaho to merge and create a single csv.

Working on Pentaho:

Here are the steps to work on Pentaho:

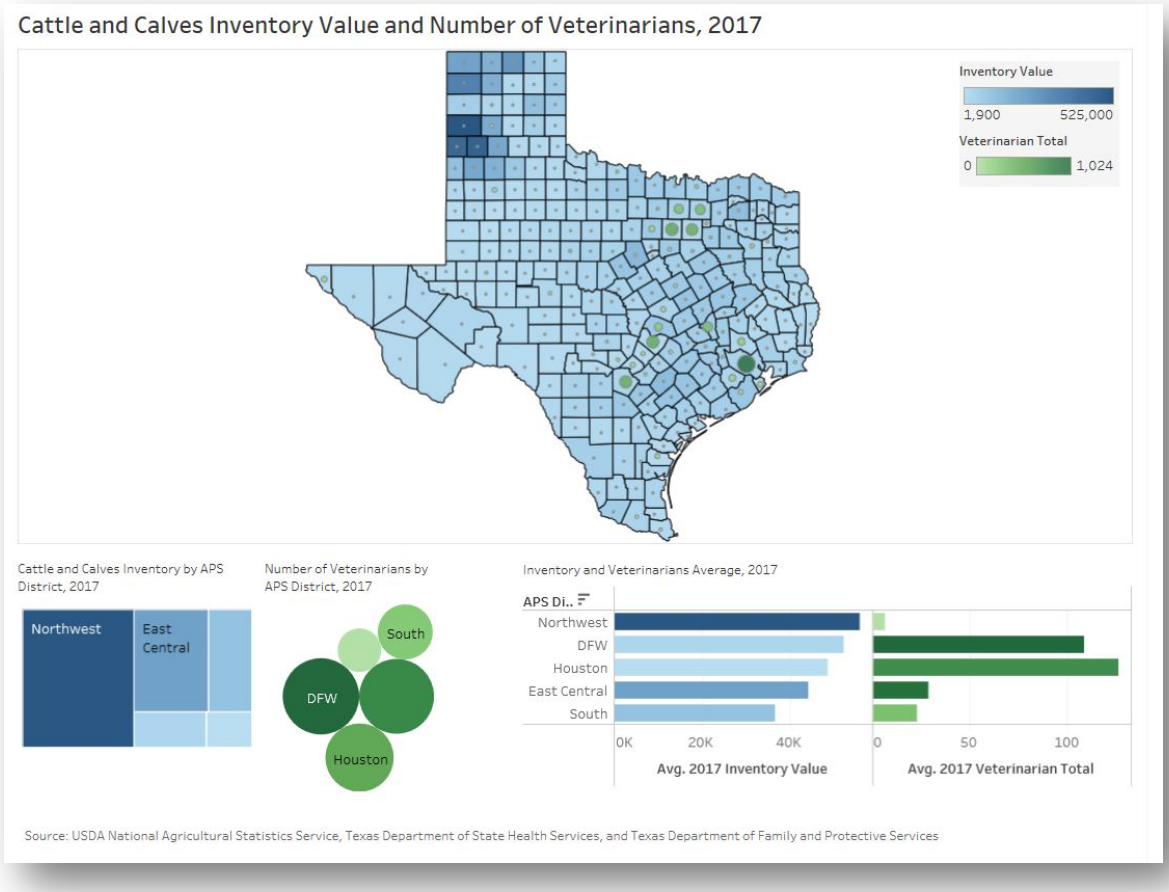
Upon opening the Pentaho file named "**Final_Project_Test.ktr**", all you have to do is just "run" the schema you see. Everything it's already done: cleaning, sorts, and merges. After running that you should

expect to find no errors during the run processing and get one single csv file output named "Merge_3_Files.csv". This is the file used on Tableau to create the visualization of this merged dataset. Please note that you may have to download and install Python's library called "selenium" and Chromedriver.

Working on Tableau:

The Tableau file you'll have to look for is named as "**Final_Project_VilasBoas.twb**". The "Merge_3_Files.csv" we've created in Pentaho it's already imported into Tableau and you should be able to see it as a dataset on the "Data Source" tab. By navigating on Tableau you'll also see other tables with the visualizations, however the main tab you should look at is the "Final Dashboard" which the ultimate visualization for the project. You'll notice that I'm not using all tabs I've created in the final dashboard.

Here is how the Final Dashboard should look like:



END.